面向在线社交网络用户生成内容的饮食话题发现研究*

张晓勇[1,2] 周清清[1,2] 章成志[1,2,3]
[1](南京理工大学经济管理学院 南京 210094)
[2](杭州师范大学阿里巴巴复杂科学研究中心 杭州 311121)
[3](江苏省数据工程与知识服务重点实验室(南京大学) 南京 210093)

摘要: 【目的】通过大规模文本聚类技术进行话题检测，并自动拣选优质话题。【方法】以新浪微博上与饮食相关的微博内容为数据源，结合文本聚类与深度学习知识进行话题检测。通过匹配微博发布的月份，将微博划分为四季微博；使用向量空间模型和文本聚类方法，对不同季节的微博进行话题检测，获得候选话题；结合深度学习知识，提出主题覆盖率概念，用以自动评价话题质量，去除低质量话题。【结果】基于主题覆盖率的话题筛选结果符合人工拣选预期，抽取获得主题覆盖率高于 0.5 的优质话题。【局限】话题检测质量的评价主要以定性评价为主。【结论】通过计算主题覆盖率来自动选择优质话题，该方法效率高，通用性强，获得的话题便于理解，较好地揭示了四季中饮食微博的话题分布。

关键词: 话题检测 用户生成内容 主题覆盖率 饮食挖掘
分类号: G353

1 引 言

Web2.0 理念和技术的发展，带动了社交媒体的迅速发展。多种多样的社交平台，为用户之间的交流提供了极大的便捷。越来越多的人通过社交网络分享自己对事物的观点。与此同时，随着生活水平的提高，人们对饮食的关注日益增加，人们在社交网络上分享美食、推荐菜谱、探讨饮食功效、寻找地方特色饮食。微博作为用户获取和分享信息的主要平台，存在大量有关饮食的评论内容。据统计，截至 2015 年 12 月，我国新浪微博用户规模达 2.3 亿，其中有 36.7%的用户通过微博分享周边美食、景点[1]。因此，基于微博数据进行饮食话题检测具有可行性与可靠性。

社交网络的快速普及和网民参与热情的空前高涨，导致网络信息的爆炸增长[2]，如何从繁杂、海量、异构的社交网络评论中高效而准确地定位热点话题，早已成为舆情监控、竞争情报等领域的研究热点[3-5]。传统的话题检测主要针对普通文本，通过大规模文本聚类获得话题[6]。在这种技术下，话题一般用代表该话题的类簇内的所有文档来表示，只包含文档的类别信息，不便于理解，往往需要通过人工审核来确定优质话题。

本文以新浪微博为研究对象，结合文本聚类与深度学习知识进行话题检测，实现优质话题的自动拣选。在文本表示模型中，结合微博语料特征筛选特征词，从而解决数据稀疏问题，提升聚类效率。在聚类过程中，使用 K-means 算法对微博进行聚类，并根据聚类评估结果确定类簇总数，获得候选话题。通过计算

通讯作者：章成志，ORCID：0000-0001-8121-4796，E-mail：zhangcz@njust.edu.cn。

*本文系国家社会科学基金项目“在线社交网络中基于用户的知识组织模式研究”(项目编号：14BTQ033)、国家社会科学基金重点项目“大数据环境下社会舆情与决策支持方法体系研究”(项目编号：14AZD084)和江苏省普通高校研究生科研创新(实践)计划项目“基于社交媒体的多粒度电影评论挖掘研究”(项目编号：SJLX15_0166)的研究成果之一。

主题覆盖率自动评价话题质量，去除低质量话题，避免人工拣选优质话题的步骤，提高了话题检测效率。

2 相关工作概述

2.1 话题检测与跟踪相关研究

互联网的飞速发展导致信息资源的高速增长，如何高效检索网络中的热点话题，已成为舆情监控、竞争情报等领域的热点[3]。话题检测与跟踪(Topic Detection and Tracking, TDT)技术就是在这种情况下应运而生的。该技术旨在解决信息过载问题[7]，自动地将相关话题的信息汇总，以供人查阅[8-9]。目前，TDT 的研究对象集中在网络新闻报道和博客上，关注点多为报道切分、话题跟踪、话题发现和新事件发现等[8]。

传统的话题发现技术主要使用聚类方法，常用的有: K-means 算法[10-11]、层次聚类法[12]、中心向量法[7, 13]、Single-Pass[13-14]等。这些方法在普通文本的话题检测任务中取得了很好的效果，如在 TDT 语料中进行的话题检测任务[15]。这种技术通常使用类簇内的所有文档来表示话题，不便于理解，往往需要通过人工审核获得优质话题。此外，随着话题模型的兴起[16-18]，一些研究通过 LDA 模型[17]及其扩展模型获取话题[19]。如文献[20]基于 LDA 话题模型抽取科技文献的话题，然后计算话题的强度和影响力，并基于此进行趋势分析；文献[21]结合 LDA 模型和仿射传播的自适应聚类算法实现话题发现；文献[22]考虑微博联系人关联关系和文本关联关系，提出一种适合微博主题挖掘的 MB-LDA 模型。这种技术的缺点在于抽取的主题词可解释性较差，且时间成本较高。

除以上提及的几种具有代表性的技术外，还有许多各具特色的话题发现技术。这些技术各有优势，目前还没有统一的评价标准。故而在实际应用过程中需要针对具体的需求进行选择。本文综合对比多种算法进行话题抽取，并结合深度学习知识，提出一种称为“主题覆盖率”的指标，用来自动评价话题质量，提高了话题检测的效率。

2.2 饮食挖掘相关研究

目前饮食挖掘研究多集中在史学[23-26]、社会学[27-29]、地理学[30-31]等领域中，旨在研究饮食文化变革对这些领域产生的影响。由于缺乏系统的数据支持，相关研究多通过实地考察、分析史料等定性化的途径进行[32]，定量和系统化的研究较少。

随着互联网上菜谱数据的日渐丰富，饮食挖掘领域开始出现一些定量化的研究工作。文献[33]通过分析多个国家和地区的 56 498 份菜谱数据，证明西方烹饪倾向于使用多种香料形成多种口味混合，比较满足所谓食物配对假设(Food Pairing Hypothesis)，而东方饮食则相反。文献[34]通过分析小规模菜谱，认为气候是影响厨师调味品选择的主要因素；而文献[35]则通过统计分析中国 20 个菜系共 8 498 份菜谱，证明地理距离比气候对饮食习惯的影响更大。

综上所述，依托于互联网提供的菜谱数据及社交网络中的评论信息，饮食领域定量化的研究成为可能。现有的饮食挖掘多集中于菜谱数据: 分析地理距离、气候等对饮食偏好的影响，探索不同地区食材搭配的偏好等，基于饮食评论的话题发现研究则较少。本文使用向量空间模型和文本聚类方法，获得饮食评论中的相关话题；结合深度学习知识，通过计算主题覆盖率自动拣选优质话题，提高话题检测效率。同时，实验结果有效地揭示了微博中饮食话题的分布特点，有助于进一步挖掘消费者在饮食领域的关注及需求。

3 研究框架与关键技术描述

3.1 研究框架

为了从海量数据中挖掘人们感兴趣的饮食话题，本文以新浪微博内容为研究对象，进行饮食话题的发现工作。由于在不同季节，饮食话题的分布差异较大，故针对不同季节的微博分别进行话题检测。首先，从新浪微博上采集与饮食相关的微博，并依据发布月份划分为四季微博；其次，基于文本表示模型及文本聚类获得话题；最后结合深度学习知识，基于主题覆盖率拣选优质话题。具体研究框架如图 1 所示。

3.2 微博内容表示模型及特征筛选

本文采用向量空间模型表示饮食微博内容，并结合微博语料特征筛选特征项。

(1) 文本预处理

在预处理部分，使用 OPENCC①对微博正文进行繁简转化，通过结巴中文分词②完成分词与词性标注。由于饮食微博中存在大量的菜名，故将菜名数据加入到结巴的自定义词典中进行分词。

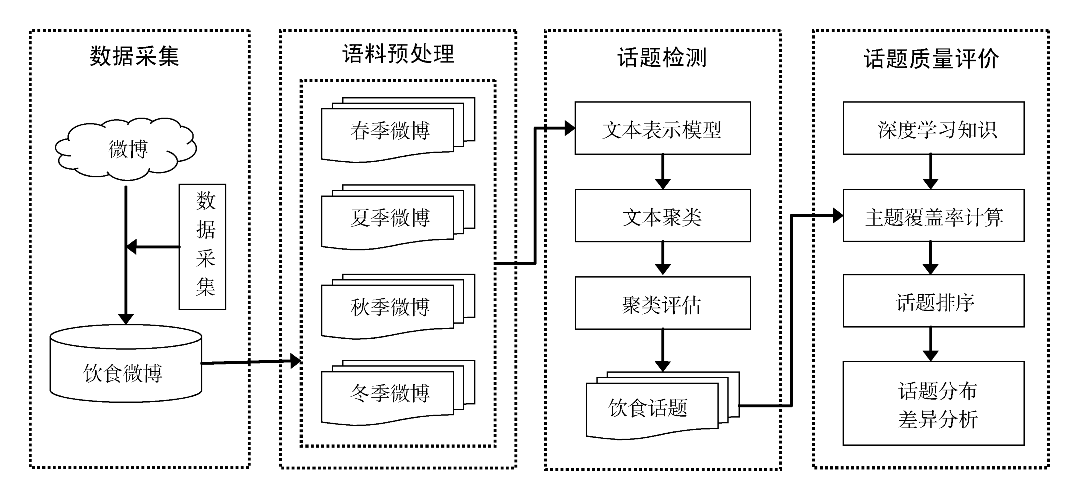


图 1　基于用户生成内容的饮食话题发现框架

(2) 向量空间模型

向量空间模型[36](Vector Space Model, VSM)由 Salton 等于 1973 年提出，其核心思想是将文本表示为文档空间的向量，把从文本筛选出的一个特征词条作为文本的一维。假设文本空间的特征项总数为 M，则第 i 个文本 d_i 可以表示为:

$$V(d_i)=(f_1,w_1(d_1);f_2,w_2(d_2);\cdots f_M,w_M(d_M)) \qquad (1)$$

其中，f_j 为第 j 项特征；w_j 为特征 f_j 在文本 d_i 中的权重，本文采用 tf-idf 算法获得其权重，公式如下:

$$W_j(d)=\frac{tf_j(d)\cdot\log(N/n_j)}{\sqrt{\sum_j(tf_j(d)\cdot\log(N/n_j))^2}} \qquad (2)$$

其中，$tf_j(d)$ 为特征 f_j 在文档 d 中的词频，n_j 为语料库中包含词 f_j 的文档总数，即通常所说的文档频率(DF 值)，N 为语料库中的文档总数。

(3) 特征项过滤策略

因语料规模较大，本文以单个词作为向量空间的特征项。在分词并过滤所有停用词后，微博短文本中仍存在大量如表情符、语气词等与话题挖掘无关的高频词。因此，需要先过滤微博中所有的表情符；通过统计特征词的文档频率(DF)，过滤掉 DF 最高的前 100 个高频词项以及 DF 值低于 100 的低频词项(两个阈值均通过人工核准确定)。需要过滤的特征项如表 1 所示:

表 1　无效特征项示例

类别	词项示例
表情符	[鼓掌]; [挖鼻]; [鄙视]; [心]; [黑线]…
DF(Top-100)	吃(2432458); 位置(1169031); 做(863060); 想(686173); 爱(65713); 说(528361); 中(429700)…
DF<100	黑伤(20); 手瓜(40); 甘长(60); 铁马(80); 歌德(99)…

从表 1 可以看出，这些高频词和表情符在大部分微博中都出现，区分性不强；低频词大多为无意义的用户昵称，话题相关性不强，故都可以过滤。

3.3　文本聚类

本文数据量较大，需要对约 500 万条饮食微博进行聚类。考虑到时间成本和话题抽取的可解释性，综合对比多种算法，最终选择运行速度最快，且话题可解释性也最好的 K-means 算法[37]对文本进行聚类。K-means 算法是一种基于原型(本文为类簇质心)的聚类技术，质心即类簇中心点。该算法随机选择 K 个初始质心，其中 K 为用户指定的类簇总数；计算每个点与质心之间的欧几里得距离，将每个点指派到距离最近的质心，而指派到一个质心的点集为一个簇，根据簇内的点，更新每个簇的质心；重复指派和更新步骤，直到质心不发生变化，则完成聚类。

由于 K-means 算法需要指定类簇总数，故本文指

①http://opencc.byvoid.com.

②http://www.oschina.net/p/jieba.

定类簇数 K=10、15…45、50，分别进行聚类，并根据聚类评估结果确定类簇数。

3.4　主题覆盖率

为避免传统方法中，通过人工拣选确定优质话题的步骤，结合深度学习知识和词语相似度计算，提出一种称为“主题覆盖率”的指标，用以评价话题质量。以下对主题覆盖率和词语相似度计算等关键技术及概念进行描述。

(1) 主题覆盖率计算

为定量评价不同话题的质量，参考文献[38]提出的“类内凝聚度”概念，并结合深度学习知识进行扩展，提出“主题覆盖率”概念。

文献[38]定义了两个概念：核心代表特征和核心文章。其中，核心代表特征是指在聚类结果中，某一类簇下 DF 值最高的 20 个特征；核心文章是指包含 m 以上个核心代表特征的文章。最终定义类内凝聚度 $\varepsilon = c_i / N$，c_i 表示核心文章总数，N 表示类簇内的文章总数。

由类内凝聚度概念可知，核心文章仅依据统计特征获得，与核心代表特征之间没有语义上的关联。由于微博的短文本特性及数据稀疏性，即使 m 值为 1，能够达到标准的微博占比也极小。

为获得核心代表特征与微博之间语义上的联系，本文结合深度学习知识，提出“主题覆盖率”概念：将核心代表特征定义为某一类簇下 DF 值最高的前 n 个特征，记作 Top-n；将核心微博定义为至少有 m 个词项与核心代表特征的词语相似度大于 p 的微博。取 Top-n=20, m=3, p=0.9，则主题覆盖率计算公式如下：

$$\gamma = c_i / N \qquad (3)$$

其中，c_i 表示核心微博总数，N 表示类簇内的微博总数。主题覆盖率值域为[0,1]，值越大，表明可以用核心代表特征表示的微博总数越多，即主题越显著，话题质量越优。

(2) 词语相似度计算

在计算主题覆盖率的过程中，需要通过计算特征词与核心代表特征之间的相似度确认核心微博总数。为计算词语相似度，本文基于深度学习知识，使用 Hinton[39]提出的 Distribute Representation 方法表征词向量，旨在将词项表达为维数较低而且固定的实数向量，通过向量空间上的相似度表示文本语义上的相似度。由于该方法更适用于大规模的计算，近年来得到广泛应用。

为使用上述方法，本文利用 Word2Vec①中的 Skip-Gram 模型进行文本表示，在分词后的全微博语料上训练，将词语转化为 400 维度的实数向量。由于 Cosine 距离常被用来衡量两个个体之间差异的大小，因此通过计算词向量之间的 Cosine 距离，可以衡量词语之间的相似度。Cosine 距离的值域为[–1, 1]，值越大表明词语越相似。

4　实验与结果分析

4.1　实验数据集

菜色名称数据来自美食杰网站②，由 Zhu 等[35]于 2012 年 4 月采集。该数据集涵盖中国 20 个菜系，共有 8 498 道菜肴名称。

本文的饮食微博数据来自新浪微博，定义正文中出现上述菜色名称的微博为“饮食微博”，采集新浪微博中 2013 年全年的饮食微博正文及用户基本信息，共计 8 747 190 条。其中，微博正文内容包括用户 ID、微博正文和发布时间，如表 2 所示：

表 2　微博正文内容示例

用户 ID	微博正文	发布时间
1785######	电池一碗辣酱面+闷蹄+酱蛋+一两小笼+一块炸猪排=撑坏了	2013-02-21 20: 04: 37
1700######	炸猪排配辣酱油简直太赞了！	2013-02-21 19: 59: 24

用户基本信息包含用户 ID、用户性别及用户所在地区，如表 3 所示：

表 3　用户基本信息示例

用户 ID	用户昵称	性别	所在地区
1000######	娜儿###	女	新疆乌鲁木齐
1000######	小瑞琪###	女	安徽宣城

①https://code.google.com/p/word2vec/.
②http://www.meishij.net.

基于用户 ID，将微博正文与用户基本信息相匹配，过滤掉丢失用户基本信息的微博正文后，最终获得 8 737 464 条微博。鉴于不同季节中，饮食话题的分布差异较大，依据微博的发布月份划分四季微博，进而检测不同季节的话题分布。

4.2 实验结果分析

(1) 饮食的季节微博划分

依据 2013 年农历中立春、立夏、立秋、立冬四个节气所在的公历月份，规定 2013 年 2-4 月份为春季，5-7 月为夏季，8-10 月为秋季，1 月及上一年 11 月、12 月为冬季。通过匹配微博数据集中微博的发布月份，除去丢失月份信息的 19 678 条微博，得到 2013 年各个季节的饮食微博共 8 717 786 条。

在对四季微博分别进行聚类的过程中发现，微博数据集中存在大量文本过短，不包含话题信息的“垃圾微博”。这些微博数量巨大，严重影响聚类效率和聚类结果的可解释性。本文在筛选微博特征词后，过滤掉特征词数目低于 10 的 3 783 652 条微博(此参数为经验数据)，大大改善了聚类结果的可解释性。最终得到 2013 年各个季节的有效饮食微博共计 4 934 134 条。过滤前后的四季饮食微博总数如图 2 所示：

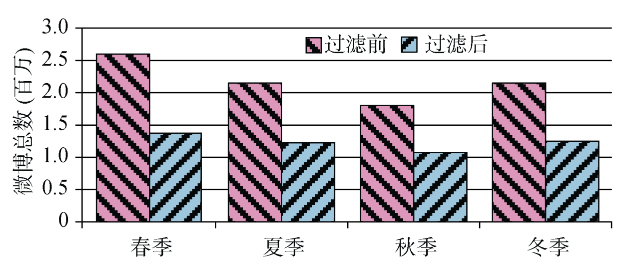


图 2 过滤前后四季微博总数

(2) 聚类与聚类评估

本文使用 K-means 算法对各个季节的微博分别进行聚类。考虑到话题检测的实际需要，指定聚类个数在 10-50 之间。因不同季节的话题总数并不一致，故指定类簇数 K=10、15…45、50，分别对各个季节的饮食微博进行 K-means 聚类，依据聚类评估结果确定最终的类簇数。

为量化评估聚类效果，使用凝聚度、轮廓系数作为有效性函数。K-means 算法是一种基于原型(本文为类簇中心点)的聚类技术，故定义簇的凝聚度(SSE)为关于簇原型的邻近度的和[40]；为衡量空间个点的绝对距离，邻近度 dist() 一般通过欧几里得距离度量。计算公式如下[40]：

$$\text{ClusterSSE} = \sum_{x \in C_i} \text{dist}(c_i, x)^2 \tag{4}$$

其中，x 代表对象，C_i 代表第 i 个簇，c_i 代表簇 C_i 的中心。凝聚度越低，表示类簇内各个对象之间的平均距离越小，簇内的凝聚性越好。

轮廓系数综合了凝聚度和分离度的优点，个体点的轮廓系数计算方法如下[40]：

①对于第 i 个对象，计算 i 到簇中所有其他对象的平均欧式距离，记为 a_i；

②对于第 i 个对象和不包含该对象的任意簇，计算该对象到给定簇中所有对象的平均欧式距离并找出最小值，该值记为 b_i；

③对于第 i 个对象，轮廓系数计算公式如下：

$$s_i = (b_i - a_i) / \max(a_i, b_i) \tag{5}$$

轮廓系数的值在–1 和 1 之间变化，值越大表明聚类质量越好。通过计算所有对象的平均轮廓系数，可以得到聚类优良性的总度量。

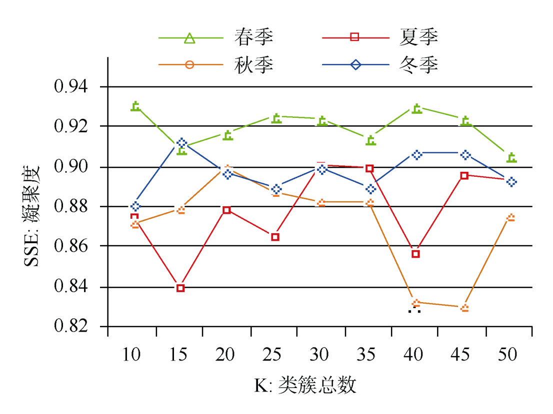


图 3 凝聚度分布图

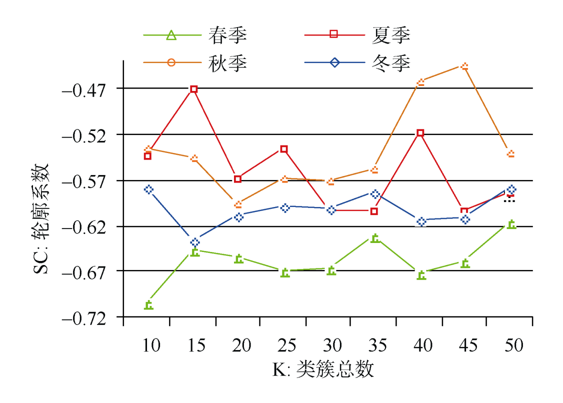


图 4 轮廓系数分布图

图 3 和图 4 为不同季节在指定不同聚类数目时的凝聚度与轮廓系数。凝聚度越低，轮廓系数越高，则聚类效果越好。可以看到，两种评价指标与类簇个数变化趋势基本一致。依据凝聚度和轮廓系数分布趋势，可以看出：春季的最优类簇数目为 50，夏季为 15，秋季为 45，冬季为 10。

(3) 基于主题覆盖率的优质话题拣选

传统的基于文本聚类的话题检测技术，在得到聚类结果后，一般通过人工拣选的方式得到主题显著性较高的类簇。这种做法人工参与程度较高，效率较低。本文通过计算每个类簇的主题覆盖率，对其主题显著度自动打分排序，通过量化评价的方式规避这一问题，可大大提高话题检测的效率。为证明该方法的有效性，表 4 以春季的聚类结果为例，对具有不同类内凝聚度的类簇进行排序。每个类簇用类内 DF 值最高的前 15 个核心代表特征表示。

表 4 春季聚类结果示例

主题覆盖率	类内凝聚度	话题名称	核心代表特征
1.0	0.99	喉咙肿痛	火#饮#水泡#咽喉#头发#蜜枣核桃#干枯#黄瓜#生姜#淡盐水#干裂#猕猴桃#嘴唇#喉干#肿痛
1.0	0.95	春节	顺利#事事#花开富贵#金银满屋#龙马精神#一帆风顺#身体健康#万事如意#百无禁忌#财源
0.98	0.29	食谱分享	倒入#翻炒#少许#洗净#料酒#均匀#烧热#捞出#葱#淀粉#小火#拌匀#生抽#酱油
0.77	0.16	养颜	皮肤#蜂蜜#姜汤#枸杞#肌肤#改善#茶#美容#牛奶#养颜#上火#醋#疼痛#火食#功效
0.65	0.10	食材：南瓜	南瓜#洗净#去皮#煸炒#泥#面团#倒入#南瓜片#南瓜饼#切片#南瓜粥#小南瓜#腌制#白糖#适量
0.56	0.12	养生粥	红枣#桂圆#枸杞#洗净#银耳#莲子#润肺#皇上#山药#百合#枸杞粥#养颜#健脾#小火#核桃
0.43	0.029	无法确认	食物#酒后#脂肪#食品#健康#牛奶#饮食#水果#蔬菜#维生素#春季#作用#香蕉
0.31	0.025	无法确认	太阳#微风#晒#阳光#天气#心情#幸福#走#月亮#下午#干杯#早上#咖啡#花
0.28	0.00025	生病	生病#一家#吐#开#可怜#餐厅#找#悲伤#卖#晕#老板#特别#走#食#三口
0.02	0.0033	无法确认	走#幸福#朋友#可爱#送#特别#时间#中国#包#笑#死#同#开#昨天#找

表 4 中部分话题被标记为“无法确认”，是由于其核心代表特征之间相关性不强，人工甄别无法确定其主题。实验结果表明主题覆盖率较高(>0.5)的类簇主题显著性较高，而主题覆盖率过低的类簇则很难判断其主题。这种基于主题覆盖率的排序符合人工拣选的预期，很好地解释了四季有关饮食的话题分布状况，同时也证实本文方法可行。

表 4 比较不同质量话题下，类内凝聚度及主题覆盖率的取值状况。可以看出：增加语义关联后的主题覆盖率取值更为合理，分布更为均匀，如“食谱分享”、“养颜”话题；由于数据稀疏问题，单纯依靠统计特征获得的核心微博数量过低，导致类内凝聚度在大部分话题下的取值都很低，无法有效评价话题质量，如“生病”及其他“无法确认”的话题。

(4) 对比实验

为论证方法的有效性，增加两组对比实验：一组基于 Doc Embedding 模型[41]，结合 K-means 聚类获得话题；一组基于 LDA 话题生成模型[17]获得话题。以春季为例，两组实验都指定话题总数为 50，在春季微博上进行话题检测，相关结果示例分别如表 5 和表 6 所示：

表 5 基于 Doc Embedding 技术获得的春季话题示例

序号	主题覆盖率	核心代表特征(Top_n=10)
1	0.78	洗净#倒入#翻炒#少许#料酒#捞出#小火#切成#生抽#均匀
2	0.64	套餐#价值#享#团购#份#售#原价#选#今日#通用
3	0.53	100#适量#50#30#20#材料#原料#洗净#牛奶#面粉
4	0.49	倒入#料酒#翻炒#少许#小火#淀粉#酱油#均匀#捞出#生抽
5	0.37	食物#健康#蜂蜜#火#皮肤#饮#功效#头发#脂肪#饮食
6	0.30	鱿鱼#酸辣粉#土豆#烤肉#汉堡#肉夹馍#小丸子#炸#章鱼#披萨
7	0.30	苹果#午餐#牛奶#一杯#米饭#香蕉#水果#饮食#早上#豆浆
8	0.27	生日#谢谢#生日快乐#礼物#送#祝#快乐#亲爱#可爱#感谢
9	0.27	菜谱#一道#豆果#网#厨房#天下#收藏#看吧#简单#大全
10	0.22	红枣#煲#炖#百合#洗净#枸杞#山药#桂圆#银耳#莲子

表 6 基于 LDA 模型获得的春季话题示例

序号	主题词(Top_n=10)
1	山东#阿姨#小小#糖醋#蟹黄#老抽#无比#牛肉面#丁丁#济南
2	食堂#特产#老爸#鸡丁#公司#江南#感受#只能#零食#来到
3	大道#素食#微信#稀饭#师傅#旅行#飞机#吃饱#大雨#大人
4	成都#价值#伤心#晚安#飞吻#无敌#团购#妹子#起床#等待
5	肉夹馍#米线#炒面#土豆#面筋#鲜虾#蟹块#菠菜#清炒#豆腐
6	热干面#蜜糖#年糕#诱惑#肯德基#绿茶#爆炒#中路#扬州#栗子
7	纳西#记录#排骨#蚕豆#肉片#风味#鸡丝#日式#莴苣#厨艺
8	取代#尼玛#烧饼#西路#电影#领取#预定#满意#面食#进口
9	武汉#虾仁#花椒#黄瓜#精选#打包#西安#红豆#凉拌#山东
10	寿司#荠菜#肉丝#荞麦#香干#鲤鱼#酸菜#西湖#鲍鱼#玫瑰

通过 Doc Embedding 技术，将每段微博正文表达为 100 维的向量；通过 K-means 算法对微博进行聚类，获得相关话题；基于主题覆盖率，对话题进行排序，相关话题用 10 个核心代表特征表示。从表 5 可以看出：与表 4 中的春季话题相比，该方法获得的话题可解释性较差，话题种类较为单一。另外，基于主题覆盖率的话题质量排序符合人工拣选的预期，再次证明该指标的有效性。

依据词性和统计特征，过滤掉与主题无关的词项；通过 LDA 主题建模获得话题分布。每个话题通过最能代表该主题的 10 个主题词表示。从表 6 可以看出：该方法获得的话题难以解释，且话题与话题之间的区分度也很低。主题词的构成模式基本为“位置+人物+食物或食材”，如话题 1 中的“山东+阿姨+蟹黄”，话题 2 中的“食堂+老爸+特产”等。

通过以上对比实验可看出：本文结合向量空间模型及文本聚类技术的话题检测方法，获得的话题可解释性更强，各个季节的话题分布也符合实际状况；基于主题凝聚度的话题质量评价方法效率高，通用性强，可以替代人工拣选高质量话题的步骤。

(5) 话题分布差异分析

为衡量各个季节内话题分布的状况，图 5 给出各个季节的主题覆盖率分布状况。

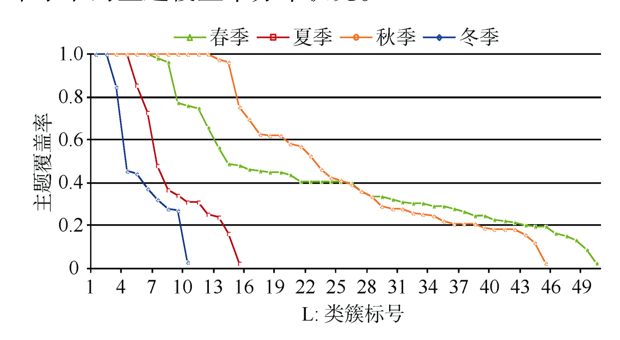


图 5 类内凝聚度分布

由图 5 可以看出，春、秋两季话题总数较多，并且主题覆盖率高的话题数目也远高于夏、冬两季。为此，以各个季节中主题覆盖率高于 0.4 且具有代表性的话题为例，将各个话题分为几个大类，对比分析四季的话题分布差异及其原因。经人工审核，本文将各个季节的饮食话题归纳为“功效”、“节日”、“烹饪”、“旅行”四个大主题，在每个大主题下枚举当季的相关话题，话题示例如表 7 所示：

表 7 四季代表性话题示例

主题	春季代表性话题	秋季代表性话题
功效	止咳#咳嗽#蜂蜜#萝卜#风寒#白萝卜#姜枣汤#伤风#鲜梨#祛痰 火#饮#水泡#咽喉#头发#蜜枣核桃#干枯#黄瓜#生姜#淡盐水 皮肤#蜂蜜#姜汤#枸杞#肌肤#改善#茶#美容#牛奶#养颜 红枣#桂圆#枸杞#洗净#银耳#莲子#润肺#皇上#山药#百合	止咳#咳嗽#蜂蜜#萝卜#风寒#伤风#姜枣汤#白萝卜#鲜梨#肺痈 火#水泡#咽喉#头发#黄瓜#干裂#猕猴桃#嘴唇#生姜#淡盐水 生津#化痰#消暑#减肥#健康#养颜#银耳#冬瓜汤#止咳#瘦身 红枣#秋季#养生#蜂蜜#食物#枸杞#功效#润肺#百合#滋阴#
节日	顺利#事事#花开富贵#金银满屋#龙马精神#一帆风顺#身体健康#万事如意#百无禁忌#财源	月饼#五仁#蛋黄#蛋黄酥#中秋节#鲜肉#莲蓉#豆沙#广式#馅
烹饪	洗净#料酒#小火#适量#少许#捞出#拌匀#生抽#倒入#淀粉	洗净#适量#倒入#少许#小火#入#料酒#拌匀#捞出#切成
旅行	北京#爆肚#豆汁#炒肝#炸酱面#小吃#成都#担担面#汤圆#中国 台湾#小吃#夜市#凤梨酥#包#台北#卤肉饭#大肠#牛轧糖#小肠	旅行#探索#旅程#感受#世界#旅途#欣赏#梦想#文化#美景 台湾#夜市#小吃#凤梨酥#包#大肠#小肠#士林#台北#面线
主题	**夏季代表性话题**	**冬季代表性话题**
功效	生津#化痰#消暑#百合#健康#润肺#养颜#银耳#冬瓜汤#火 火#饮#水泡#咽喉#肿痛#蜜枣核桃#生姜#头发#梨#猕猴桃 MM#试#少女#热荐#冰清玉洁#韩国#晶体#酵素#大 S#红嫩	止咳#咳嗽#蜂蜜#萝卜#风寒#冰糖水#伤风#姜枣汤#白萝卜#咳咽 食物#健康#牛奶#红枣#营养#养生#饮食#蜂蜜#苹果#百合
烹饪	洗净#倒入#翻炒#料酒#捞出#小火#适量#生抽#均匀#淀粉	倒入#洗净#翻炒#少许#料酒#小火#入#4.#捞出#适量

受限于篇幅，相关话题用 10 个核心代表特征表示。对比表中各个季节话题分布，可以得出以下几个结论：

①有关饮食"功效"的话题在全年都非常显著，具体到某类功效又会随着季节特点发生变化。如"降火"、"止咳"在四季都有分布；春秋两季特别的有"养颜"、"滋阴"话题；夏季增加"消暑"话题；冬季增加"养生饮食"话题。

②有关"烹饪"教程的话题在全年都有分布，且同质性较高。

③春秋两季气温适宜，用户外出游玩机会较多，故与饮食相关的话题较多，如"旅行"类中的"地方特色小吃"、"旅游景点"。

④春秋两季重要的传统节日较多，如"春节"、"中秋节"，用户倾向于在特定节日分享具有代表性的食物，如"月饼"。

⑤夏季、冬季由于气候较为极端，用户出行较少，缺少与"旅行"相关的话题；与饮食相关的节假日较少，缺少与"节假日"相关的话题。

通过以上分析，解释了春、秋两季话题数多于夏、冬两季的原因；同时也证明，基于本文方法获得的四季话题与实际状况相契合。

5 总结与展望

网络中丰富的菜谱数据和社交网络上海量的饮食评论为饮食挖掘研究提供了数据支持，如何从这些评论中检测出热点话题，进而为消费者和营销商提供决策依据，已经成为各方普遍关注的问题。传统的话题检测任务主要通过大规模文本聚类获得话题，由于该方法获得的话题只包含类别信息，不便于人们理解，往往需要人工审核去除劣质话题，话题检测效率较低。

本文以新浪微博为数据来源，结合文本聚类与深度学习知识进行话题检测。在通过文本聚类获得四季饮食话题后，基于主题覆盖率自动拣选优质话题。本文方法通用性强，效率较高，避免了在聚类完成后人工拣选话题的步骤。实验结果较好地揭示了四季中饮食微博的话题分布，有助于进一步挖掘消费者在饮食领域的关注热点。在今后的工作中，将进一步考虑以下内容：该话题检测方法在其他领域的推广；结合特征词抽取技术，更准确深入地实现话题检测任务。

参考文献：

[1] 中国互联网络信息中心.第 37 次中国互联网络发展状况统计报告[R/OL]. (2016-01-22). [2016-05-25]. http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/201601/P020160122469130059846.pdf. (China Internet Network Information Center. The 37th Report of Chinese Internet Development [R/OL]. (2016-01-22). [2016-05-25]. http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/201601/P020160122469130059846.pdf.)

[2] 殷风景，肖卫东，葛斌，等. 一种面向网络话题发现的增量文本聚类算法[J]. 计算机应用研究，2011, 28(1): 54-57. (Yin Fengjing, Xiao Weidong, Ge Bin, et al. Incremental Algorithm for Clustering Texts in Internet-oriented Topic Detection[J]. Application Research of Computers, 2011, 28(1): 54-57.)

[3] 王伟，许鑫. 基于聚类的网络舆情热点发现及分析[J]. 现代图书情报技术，2009(3): 74-79. (Wang Wei, Xu Xin. Online Public Opinion Hotspot Detection and Analysis Based on Document Clustering[J]. New Technology of Library & Information Service, 2009(3): 74-79.)

[4] 徐东亮. 基于聚类分析的网络论坛舆情信息挖掘技术研究[D]. 哈尔滨：哈尔滨工业大学，2010. (Xu Dongliang. Research of Public Opinion Information Mining on Bulletin Board Systems Based on Cluster Analysis[D]. Harbin: Harbin Institute of Technology, 2010.)

[5] 朱恒民，李青. 面向话题衍生性的微博网络舆情传播模型研究[J]. 现代图书情报技术，2012(5): 60-64. (Zhu Hengmin, Li Qing. Public Opinion Propagation Model with Topic Derivatives in the Micro-blog Network [J]. New Technology of Library & Information Service, 2012(5): 60-64.)

[6] 洪宇，张宇，刘挺，等. 话题检测与跟踪的评测及研究综述[J]. 中文信息学报，2007, 21(6): 71-87. (Hong Yu, Zhang Yu, Liu Ting, et al. Topic Detection and Tracking Review[J]. Journal of Chinese Information Processing, 2007, 21(6): 71-87.)

[7] Allan J, Carbonell J, Doddington G, et al. Topic Detection and Tracking Pilot Study Final Report[C]. In: Proceedings of the 1998 Broadcast News Transcription and Understanding Workshop. 1998.

[8] 路荣，项亮，刘明荣，等. 基于隐主题分析和文本聚类的微博客中新闻话题的发现[J]. 模式识别与人工智能，2012, 25(3): 382-387. (Lu Rong, Xiang Liang, Liu Mingrong, et al. Discovering News Topics from Microblogs Based on Hidden Topics Analysis and Text Clustering[J]. Pattern Recognition & Artificial Intelligence, 2012, 25(3): 382-387.)

[9] 骆卫华，刘群，程学旗. 话题检测与跟踪技术的发展与研究[C]. 见：全国计算语言学联合学术会议 (JSCL-2003) 论文集. 北京：清华大学出版社，2003: 560-566. (Luo Weihua, Liu Qun, Cheng Xueqi. Development and Analysis of Technology of Topic Detection and Tracking [C]. In: Proceedings of the 7th National Conference on

Computational Linguistics. Beijing: Tsinghua University Press, 2003: 560-566.)
[10] Xu J, Croft W B. Cluster-based Language Models for Distributed Retrieval [C]. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 1999.
[11] Wartena C, Brussee R. Topic Detection by Clustering Keywords [C]. In: Proceedings of the 19th International Conference on Database and Expert Systems Application. IEEE Computer Society, 2008: 54-58.
[12] Yang Y, Pierce T, Carbonell J. A Study on Retrospective and On-line Event Detection[C]. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 1998.
[13] Jia Z Y, Qing H E, Zhang H J, et al. A News Event Detection and Tracking Algorithm Based on Dynamic Evolution Model[J]. Journal of Computer Research & Development, 2004, 41(7): 1273-1280.
[14] 贾自艳, 何清, 张海俊, 等. 一种基于动态进化模型的事件探测和追踪算法[J]. 计算机研究与发展, 2004, 41(7): 1273-1280. (Jia Ziyan, He Qing, Zhang Haijun, et al. A News Event Detection and Tracking Algorithm Based on Dynamic Evolution Model [J]. Journal of Computer Research & Development, 2004, 41(7): 1273-1280.)
[15] 马彬, 洪宇, 陆剑江, 等. 基于线索树双层聚类的微博话题检测[J]. 中文信息学报, 2012, 26(6): 121-128. (Ma Bin, Hong Yu, Lu Jianjiang, et al. A Thread-based Two-stage Clustering Method of Microblog Topic Detection[J]. Journal of Chinese Information Processing, 2012, 26(6): 121-128.)
[16] Hofmann T. Probabilistic Latent Semantic Indexing[C]. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 1999.
[17] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation[J]. Journal of Machine Learning Research, 2003, 3: 993-1022.
[18] Griths T L, Steyvers M. A Probabilistic Approach to Semantic Representation [C]. In: Proceedings of the 24th Annual Conference of the Congnitive Science Society. 2002: 381-386.
[19] 单斌, 李芳. 基于 LDA 话题演化研究方法综述[J]. 中文信息学报, 2010, 24(6): 43-49. (Shan Bin, Li Fang. A Survey of Topic Evolution Based on LDA[J]. Journal of Chinese Information Processing, 2010, 24(6): 43-49.)
[20] 贺亮, 李芳. 基于话题模型的科技文献话题发现和趋势分析[J]. 中文信息学报, 2012, 26(2): 109-115. (He Liang, Li Fang. Topic Discovery and Trend Analysis in Scientific Literature Based on Topic Model [J]. Journal of Chinese Information Processing, 2012, 26(2): 109-115.)
[21] 吴永辉, 王晓龙, 丁宇新, 等. 基于主题的自适应、在线网络热点发现方法及新闻推荐系统[J]. 电子学报, 2010, 38(11): 2620-2624. (Wu Yonghui, Wang Xiaolong, Ding Yuxin, et al. Adaptive On-Line Web Topic Detection Method for Web News Recommendation System[J]. Acta Electronica Sinica, 2010, 38(11): 2620-2624.)
[22] 张晨逸, 孙建伶, 丁轶群. 基于 MB-LDA 模型的微博主题挖掘[J]. 计算机研究与发展, 2011, 48(10): 1795-1802. (Zhang Chenyi, Sun Jianling, Ding Yiqun. Topic Mining for Microblog Based on MB-LDA Model[J]. Journal of Computer Research & Development, 2011, 48(10): 1795-1802.)
[23] Civitello L. Cuisine and Culture: A History of Food and People[M]. Wiley, 2011.
[24] Tregear A. From Stilton to Vimto: Using Food History to Re-think Typical Products in Rural Development [J]. Sociologia Ruralis, 2003, 43(2): 91-107.
[25] 王仁湘. 饮食与中国文化[M]. 北京: 人民出版社, 1993. (Wang Renxiang. Diet and Chinese Culture [M]. Beijing: People's Publishing House, 1993.)
[26] 张景明. 中国北方游牧民族饮食文化研究[M]. 北京: 文物出版社, 2008. (Zhang Jingming. Chinese Nomads Food Culture[M]. Beijing: Cultural Relics Press, 2008.)
[27] Mennell S, Murcott A, Otterloo A H V. The Sociology of Food: Eating, Diet and Culture [M]. London: Sage Publications, 1992.
[28] Beardsworth A, Keil E T. Sociology on the Menu: An Invitation to the Study of Food and Society[J]. British Journal of Sociology, 2002, 49(2): 327-328.
[29] Germov J, Williams L. A Sociology of Food and Nutrition: The Social Appetite [M]. The 3rd Edition. Oxford University Press, 2008.
[30] 陈传康. 中国饮食文化的区域分化和发展趋势[J]. 地理学报, 1994, 49(3): 226-235. (Chen Chuankang. The Culture of Chinese Diet: Regional Differentiation and Developing Trends[J]. Acta Geographica Sinica, 1994, 49 (3): 226-235.)
[31] 蔡晓梅, 司徒尚纪. 中国地理学视角的饮食文化研究回顾与展望[J]. 云南地理环境研究, 2006, 18(5): 83-88. (Cai Xiaomei, Situ Shangji. A Review on the Studies of Food Culture from Geographical Perspective [J]. Yunnan Geographic Environment Research, 2006, 18(5): 83-88.)
[32] 蓝勇. 中国饮食辛辣口味的地理分布及其成因研究[J]. 地理研究, 2001, 16(5): 229-237. (Lan Yong. On The Reasons

and Distrbution of Pungent Flavour in Chinese Food and Drink [J]. Geographical Research, 2001, 16(5): 229-237.)

[33] Ahn Y Y, Ahnert S E, Bagrow J P, et al. Flavor Network and the Principles of Food Pairing [J/OL]. Scientific Reports, 2011: Article No. 196. http://www.nature.com/articles/srep00196.

[34] Sherman P W, Billing J. Darwinian Gastronomy: Why We Use Spices [J]. Bioscience, 1999, 49(6): 453.

[35] Zhu Y X, Huang J, Zhang Z K, et al. Geography and Similarity of Regional Cuisines in China [J]. PLoS One, 2013, 8(11): e79161.

[36] Salton G, Yang C S. On the Specification of Term Values in Automatic Indexing [J]. Journal of Documentation, 1973, 29(4): 351-372.

[37] Arthur D, Vassilvitskii S. K-means++: The Advantages of Careful Seeding [C]. In: Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms. 2007: 1027-1035.

[38] 彭楠赟, 王厚峰, 凌晨添. 基于层次聚类的网络新闻热点发现[A]. //中国计算语言学研究前沿进展(2009-2011)[R]. 北京: 清华大学出版社, 2011: 487-492. (Peng Nanyun, Wang Houfeng, Ling Chentian. Event Mining in On-line News Based on Hierarchical Clustering [A]. // Advances of Computational Linguistics in China [R]. Beijing: Tsinghua University Press, 2011: 487-492.)

[39] Hinton G E. Learning Distributed Representations of Concepts [C]. In: Proceedings of the 8th Annual Meeting of the Cognitive Science Society. 1986.

[40] Tan P N, Steinbach M, Kumar V, et al. Introduction to Data Mining [M]. Pearson, 2010.

[41] Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space [OL]. ArXiv: 1301. 3781.

[42] Mikolov T, Sutskever I, Chen K, et al. Distributed Representations of Words and Phrases and Their Compositionality [J]. Advances in Neural Information Processing Systems, 2013, 26: 3111-3119.

作者贡献声明:

张晓勇: 文献调研与整理, 论文起草;
周清清: 协助完成实验, 论文修订;
章成志: 提出研究思路, 讨论研究方案, 论文最终版本修订。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: riyao95@qq.com。

[1] 张晓勇, 周清清, 章成志. Seg_Kmeans.py. 四季饮食微博K-means 聚类及聚类评价算法.

[2] 张晓勇, 周清清, 章成志. Inside_Cohesion.py. 类内凝聚度计算.

[3] 张晓勇, 周清清, 章成志. train_word2vec_model.py. 词向量训练算法.

[4] 张晓勇, 周清清, 章成志. Weibo_data_initial.txt. 新浪微博原始抓取数据.

[5] 张晓勇, 周清清, 章成志. Weibo_data_seg.txt. 新浪微博分词数据.

[6] 张晓勇, 周清清, 章成志. Inside_Cohesion_sorted.txt.四季饮食话题抽取结果及类内凝聚度排序数据.

收稿日期: 2016-05-26
收修改稿日期: 2016-07-18

Identifying Food Topics from User-Generated Contents in Microblogs

Zhang Xiaoyong[1, 2] Zhou Qingqing[1, 2] Zhang Chengzhi[1, 2, 3]
[1](School of Economics and Management, Nanjing University of Science and Technology, Nanjing 210094, China)
[2] (Alibaba Research Center for Complex Sciences, Hangzhou Normal University, Hangzhou 311121, China)
[3] (Jiangsu Key Laboratory of Data Engineering and Knowledge Service (Nanjing University), Nanjing 210093, China)

Abstract: [**Objective**] This study aims to identify microblog post topics, and then automatically extract high quality ones with the help of text clustering techniques. [**Methods**] We collected food related microblog posts from Sina Weibo as raw data, then applied text clustering and deep learning techniques to detect the target topics. First, we categorized the microblog posts by the four seasons in accordance with their publishing dates. Second, we created a vector space model and used text clustering method to retrieve candidate topics. Finally, we automatically identified the quality topics with deep learning technology. [**Results**] We automatically identified the high quality topics manually found by researchers, and their topic coverage values were all higher than 0.5. [**Limitations**] We decided the topic quality based on qualitative data. [**Conclusions**] The proposed method could extract high quality topics effectively. The retrieved topics reflect the distribution of food related microblog posts in the four seasons.

Keywords: Topic detection User-Generated Contents Topic coverage Food mining

---

德克萨斯大学图书馆成为全球第一个推出开放获取政策的图书馆

德克萨斯大学图书馆为德克萨斯大学奥斯汀分校的全体工作人员制定了正式的开放获取政策。一个适度的、能吸引图书馆工作人员将期刊文章和会议论文存储到德克萨斯大学数字资源库 Texas ScholarWorks 之中的计划也于近日获得了学校的批准。

该政策仅适用于德克萨斯大学图书馆员工, 并且是非排他性的, 这意味着工作人员在提交成果到当地存储库的同时, 还可以自由地向外部出版组织提交他们的成果。这项政策立即生效, 并且不适用于此前发布或撰写的材料。

开放获取是一项国际运动, 其目标是使所有经过同行评议出版的学术成果能够免费提供给公众和全球学术界, 开放获取包含开放成果(学术出版物和馆藏)、开放数据(研究数据)和开放教育资源(开放教科书)。

德克萨斯大学图书馆馆长 Lorraine Haricombe 在来到德克萨斯大学之前, 曾在堪萨斯大学主导实施了一项以教师为主导的开放获取政策, 这是美国首个公共机构推出这样的政策。

Haricombe 解释说: “从来到德克萨斯大学的第一天起, 在图书馆采用开放获取政策就排在我工作计划列表的前几位。德克萨斯大学图书馆致力于开放的议程, 使得学术研究的成果更易于被所需要的人访问。我希望这样的政策能进一步扩散到整个大学。”

(编译自: http://www.lib.utexas.edu/d7/about/news/libraries-institute-first-formal-open-access-policy)

(本刊讯)